Uncertain Shapelet Transform: A Shapelet-based Approach for Uncertain Time Series Classification

Michael Franklin Mbouopda and Engelbert Mephu Nguifo

Laboratory of Computing, Modelling and Optimization of the Systems – LIMOS

October 11, 2019

Context: TransiXplore project

- Aims to study astronomical transient objects
- Objects are represented by their light curve
- Light curves are subject to uncertainty
- Uncertainty is explicitly expressed
- Physicists think that a shapelet-approach could work well

Contents

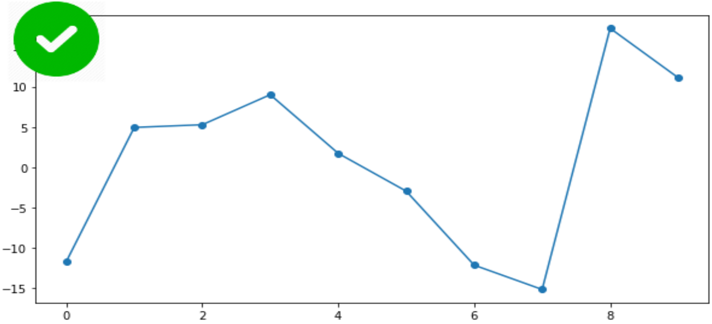
Contents

P

Introduction

- **Time series classification**: *classification* of objects modelized as *time series*. ex: Plasstic challenge
- **Time series**: sequence of chronologycal data
- Application: Physics, Medicine, Engineering, ...
- Very active field

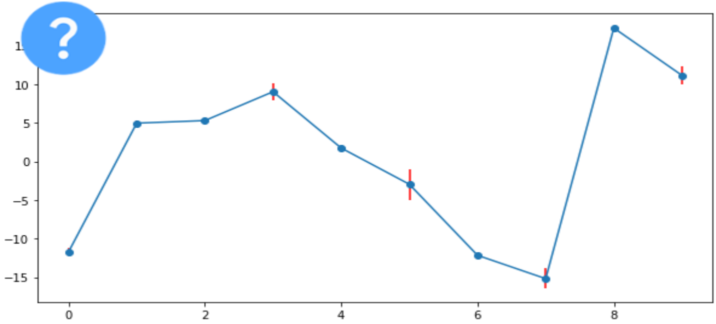
State of the art



Many classification methods

- Machine learning-based [Bagnall et al., 2017]: **Shapelet Transform**, BOOS, etc
- Deep learning-based [Fawaz et al., 2019b]: **ResNet**, FCN, etc
- Composition-based [Lines et al., 2018]: **HIVE-COTE**, FLAT-COTE

State of the art



No method found for uncertain time series, but we have

- Error analysis strategies [Taylor, 1996]
- Uncertain supervised classifiers: UDT[Tsang, Kao, Yip, Ho, and Lee, 2009]

Our motivation and method

Motivation

- Uncertainty cannot be eliminated [Taylor, 1996]
- Explicitly take uncertainty into account when it is available will lead to more accurate results

Method

1. Propagate uncertainty in shapelet transformation
2. Use an uncertain supervised classifier

Contents

Time series

Time series (without uncertainty)

Sequence of m chronological values. m is the time series length.

$$T = \{t_1, t_2, ..., t_m\}$$

Uncertain time series

A time series where each value has an uncertainty. We represent uncertainty as follow.

$$x = x_{best} \pm \delta x$$

Subsequence

Sequence of l consecutive values of a time series starting at position i

$$S = \{t_{i+1}, ..., t_{i+l}\}$$

Notion of distance

Let $S = \{s_1, s_2, ..., s_l\}$ and $R = \{r_1, r_2, ..., r_l\}$ be two subsequences of same length l.

Distance between subsequences

$$\mathrm{ED}(S, R) = \frac{1}{l}\sum_{i=1}^{l}(s_i - r_i)^2.$$

Let $T = \{t_1, t_2, ..., t_m\}$ be a time series of length m, $m > l$

Distance between subsequence and time series

$$\mathrm{ED}(S, T) = \min\{\mathrm{ED}(S, R)\ \forall R \in T\}$$

Notion of separation

Le D be a set of time series

Separator

sp is a sperator of D if it divides D in two sets D_1 and D_2 such that

$$D_1 = \{T \mid \mathrm{ED}(T, sp) \leq \epsilon, \forall T \in D\},$$

$$D_2 = \{T \mid \mathrm{ED}(T, sp) > \epsilon, \forall T \in D\}, \epsilon \in R.$$

Information Gain

It is a measurement of the quality of a separator.

$$IG(D, sp) = H(D) - (\frac{|D_1|}{|D|}H(D_1) + \frac{|D_2|}{|D|}H(D_2))$$

$H(D)$ is the entropy on the dataset D

Notion of Shapelet

Horned Lezard and turtle can be differenciated by the presence of horns. No need to examine every part of the body.

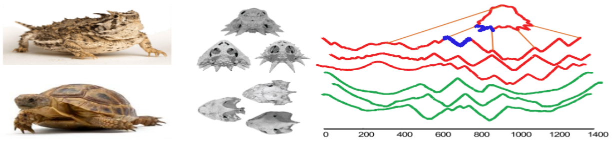


Figure 1: Horned Lizard vs Turtle.

Shapelet

It is a separator that maximizes the information gain.

$$S = \underset{sp}{\text{argmax}}(IG(D, sp))$$

Shapelet-based classification

General idea

Given a dataset of time series:

1. Select the first k shapelets
2. Compute feature vectors: vectors of distances to shapelets
3. train a classifier on the set of feature vector.

Advantages of shapelet-based classification

- Interpretability
- Robustness
- Rapid inference

Shapelet-based classification

Illustration

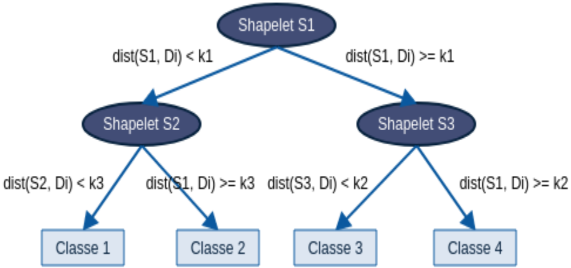


Figure 2: Shapelet decision tree illustration

Contents

Uncertainty propagation techniques

Let $x = x_{best} \pm \delta x$ and $y = y_{best} \pm \delta y$. From the book "An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements" [Taylor, 1996] we have:

Addition and substraction

$$x + y = (x_{best} + y_{best}) \pm (\delta x + \delta y)$$

$$x - y = (x_{best} - y_{best}) \pm (\delta x + \delta y)$$

Multiplication and division

$$x \cdot y = (x_{best} \cdot y_{best}) \pm (\delta x \cdot |y_{best}| + \delta y \cdot |x_{best}|)$$

$$\frac{x}{y} = (\frac{x_{best}}{y_{best}}) \pm (\frac{\delta x \cdot |y_{best}| + \delta y \cdot |x_{best}|}{|y_{best}|^2})$$

Uncertain Euclidean Distance

Let $S \pm \delta S = < s_1 \pm \delta s_1, s_2 \pm \delta s_2, ..., s_l \pm \delta s_l >$ and $R \pm \delta R = < r_1 \pm \delta r_1, r_2 \pm \delta r_2, ..., r_l \pm \delta r_l >$ be two uncertain subsequences. If they were not uncertain, then ST algorithm would compute the distance between them as follow:

$$ED(S, R) = \frac{1}{l} \sum_{i=1}^{l} (s_i - r_i)^2$$

Using the previous propagation techniques we define **UED** as follow:

$$UED(S \pm \delta S, R \pm \delta R) = (\frac{1}{l} \sum_{i=1}^{l} (s_i - r_i)^2) \pm (\frac{2}{l} \sum_{i=1}^{l} |s_i - r_i| \times (\delta s_i + \delta r_i))$$

$$UED(S \pm \delta S, R \pm \delta R) = ED(S, R) \pm (\frac{2}{l} \sum_{i=1}^{l} |s_i - r_i| \times (\delta s_i + \delta r_i))$$

Ordering uncertain measures

Let $x = x_{best} \pm \delta x$ and $y = y_{best} \pm \delta y$, we have the following properties:

- $x = y$ if and only if $x_{best} = y_{best}$ and $\delta x = \delta y$
- $x < y$ if and only if one of the following conditions is satisfied:
  - $x_{best} < y_{best}$
  - $x_{best} = y_{best}$ and $\frac{\delta x}{x_{best}} < \frac{\delta y}{y_{best}}$

We can now define the uncertain distance between an uncertain time series $T \pm \delta T$ and a subsequence $S \pm \delta S$

Definition

Uncertain distance between time series and subsequence

$$UED(S \pm \delta S, T \pm \delta T) = \min\{UED(S \pm \delta S, R \pm \delta R)\ \forall R \pm \delta R \in T \pm \delta T\}$$

Uncertain data classification

Flat representation

- Represent each uncertain data by a flat vector
- The first half of the vector contains best guesses and the second half contains uncertainties

Illustration

T: an uncertain time series
X: $[udist_1, udist_2, ..., udist_k]$, where
$udist_j = UED(T, shapelet_j) = d_j \pm \delta d_j$, then:

$$Flat(X) = [d_1, d_2, ..., d_k, \delta d_1, \delta d_2, ..., \delta d_k]$$

Uncertain data classification

Flat Classification of uncertain time series

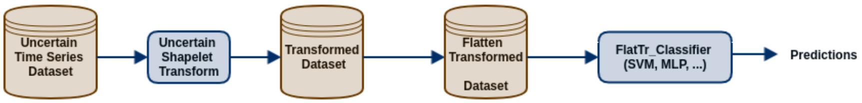


Figure 3: Flat classification archtecture

Pros

- Simplicity
- Uncertainty is taken into account through all the process

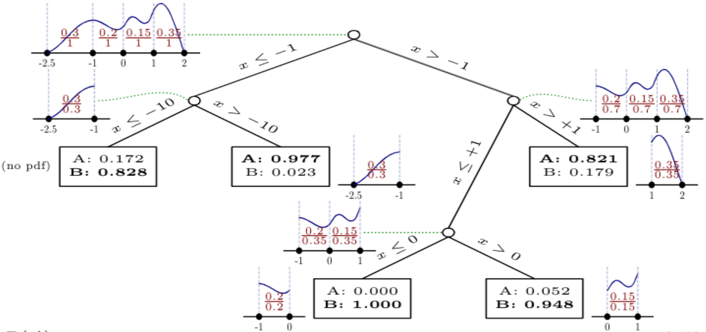
Cons

- Risk of having large flattened vectors.
- The classifier used (FlatTr_Classifier) is not aware of uncertainty

Uncertain data classification

Uncertain decision tree classifier: an Overview

- Proposed by Tsang et al. [2009]
- Awared of uncertainty in data



$$P(A) = 0 \times 0.172 + 0.3 \times 0.977 + 0.2 \times 0.0 + 0.15 \times 0.052 + 0.35 \times 0.821 = 0.59$$
$$P(B) = 0 \times 0.828 + 0.3 \times 0.023 + 0.2 \times 1.0 + 0.15 \times 0.948 + 0.35 \times 0.179 = 0.41$$

Contents

Data source

- We used 21 datasets from UEA/UCR: `http://www.timeseriesclassification.com/dataset.php`
- These datasets do not have uncertainty

Uncertain datasets are obtained by adding generated uncertainty to each dataset. The generated uncertainty follows a zero-mean normal distribution.

Source code

- Written in JAVA
- Extends the UEA/UCR time series classification source code
- Available on Github:
  https://github.com/frankl1/Uncertain-Shapelet-Transform

Results

Accuracy comparison

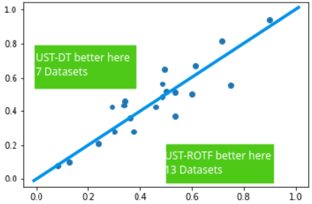


Figure 4: UST-DT vs UST-ROTF

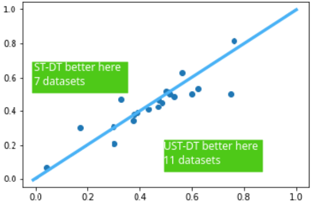


Figure 5: ST-DT vs UST-DT

- From figure 4 $\rightarrow$ Classifier matter
- From figure 5 $\rightarrow$ Handling uncertainty matter

Results

Underfitted datasets

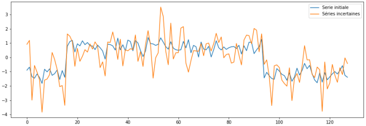


Figure 6: An instance from CBF

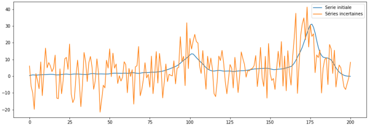


Figure 7: An instance from Fungi

- Generated uncertainty is very high
- Time series classifiers are **Very sensitive** to adversarial attack [Fawaz et al., 2019a; Karim et al., 2019]

Contents

Conclusion

What we did

- We explored how to classify uncertain time series
- We proposed Uncertain Shapelet Transformation, which shown interesting results

What we will do

- Use a supervised classifier that is aware of uncertainty, for instance UDT[Tsang et al., 2009]
- Evaluate UST on a real uncertain datasets: TransiXplore data
- Evaluate UST on the remaining 117 UEA/UCR datasets

References I

A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31 (3):606–660, 2017.

M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas. Uncertain Time-Series Similarity: Return to the Basics. *arXiv e-prints*, art. arXiv:1208.1931, Aug 2012.

H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Adversarial attacks on deep neural networks for time series classification, 2019a.

H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019b.

References II

J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401. ACM, 2014.

J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4):851–881, 2014.

F. Karim, S. Majumdar, and H. Darabi. Adversarial attacks on time series, 2019.

J. Lines, S. Taylor, and A. Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Trans. Knowl. Discov. Data*, 12(5):52:1–52:35, July 2018. ISSN 1556-4681.

References III

T. Rakthanmanon and E. Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *proceedings of the 2013 SIAM International Conference on Data Mining*, pages 668–676. SIAM, 2013.

V. S. Siyou Fotso, E. Mephu Nguifo, and P. Vaslin. Frobenius correlation based u-shapelets discovery for time series clustering. Apr. 2018. working paper or preprint.

J. R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, 2 sub edition, 1996.

S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee. Decision trees for uncertain data. *IEEE transactions on knowledge and data engineering*, 23(1):64–78, 2009.

References IV

L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.

Thanks for your attention !